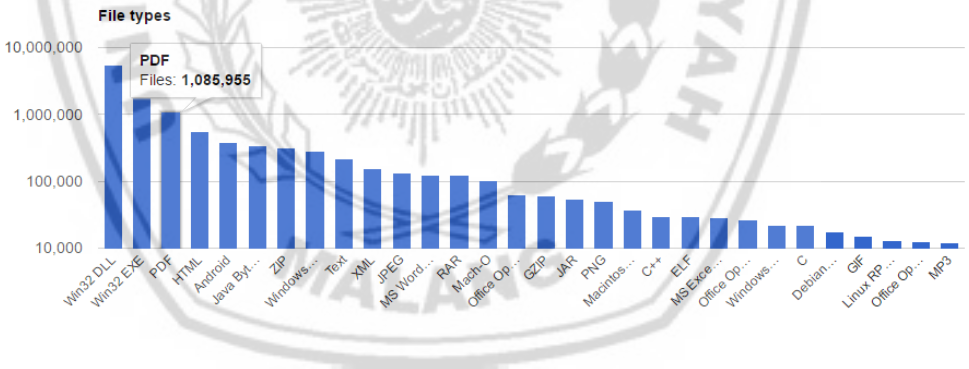
BAB I
PENDAHULUAN

1.1. Latar Belakang

Portable Document Format atau biasa disebut PDF merupakan format *file* yang sangat dibutuhkan banyak orang karena kemudahan pemakaiannya. Format PDF mendukung kemampuan untuk menyertakan *file* multimedia, akses URL langsung dan komunikasi HTTP, serta format dokumen yang multi-platform. Dengan banyaknya fitur dan didukung berbagai macam *device* dan *platform*, sehingga hal ini dapat dimanfaatkan oleh pihak yang tidak bertanggung jawab untuk menyamarkan dalam penyebaran *malware* [4].

Berdasarkan hasil statistik www.virustotal.com, PDF menduduki peringkat pertama *file* dokumen yang terjangkit *malware* dibanding *file* dokumen lain sedangkan jika dibandingkan dengan format *file* selain format *file* dokumen, PDF menempati peringkat ketiga [15]. Sebagaimana yang telah dijelaskan oleh gambar di bawah ini:



Gambar 1.1. Hasil Statistik www.virustotal.com [15] (diakses pada tanggal 23 Januari 2017)

Support Vector Machine adalah algoritma dengan teknik menentukan *Hiperplanes* terbaik untuk memisahkan antara kelas data yang sudah ditentukan. Permasalahan pertama yang harus terselesaikan dalam penelitian ini adalah bagaimana menghitung keakuratan *Support Vector Machine* sebagai teknik klasifikasi antara PDF *malware* dan yang tidak. *Support Vector Machine* adalah salah satu metode klasifikasi yang memiliki keakuratan yang tinggi yaitu nilai rata-rata *True Positive* 0,99885 dan rata-rata *False Negative* 0,00011 dengan skala

1 [13], nilai tersebut dapat berbeda dengan hasil yang akan didapat dalam penelitian ini karena perbedaan pemilihan fitur.

Random Decision Forest adalah algoritma dengan menerapkan beberapa *Decision Tree* untuk memvoting fitur-fitur sebagai klasifikasi data. Dalam metode kedua ini juga harus dihitung tingkat keakuratannya untuk dibandingkan dengan metode SVM. *Random Decision Forest* juga memiliki tingkat keakuratan yang cukup tinggi dibanding metode lain berdasarkan penelitian sebelumnya yang memiliki nilai rata-rata *True Positive* 0,9998 dan rata-rata *False Positive* 0,0208 dengan skala 1 [12], nilai tersebut juga dapat berbeda dengan hasil yang akan didapat dalam penelitian ini karena perbedaan pemilihan fitur.

Berdasarkan penelitian sebelumnya, teknik klasifikasi dengan metode SVM dan *Random Forest* memiliki tingkat keakuratan yang tinggi, menjadikan kedua metode tersebut sangat cocok untuk dibandingkan sebagai pengklasifikasian *file* PDF *malware*. Untuk membandingkan tingkat keakuratannya peneliti menggunakan perhitungan *Confusion Matrix* yang yang terdiri dari nilai rata-rata *true positive*, rata-rata *false positive*, rata-rata *true negative*, rata-rata *false negative*, *accuracy*, dan *precision.*

Tujuan dari penelitian ini adalah pengklasifikasian *file* PDF yang telah ditanamkan *malware* dan yang tidak menggunakan algoritma *Support Vector Machine* dan *Random Forest*, serta untuk mengetahui seberapa akurat antara kedua algoritma tersebut.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka dapat dirumuskan sebuah permasalahan sebagai berikut:

1. Bagaimana menerapkan perhitungan *Confusion Matrix* untuk menghitung akurasi algoritma *Support Vector Machine* dalam pengklasifikasian *file malware* PDF?
2. Bagaimana menerapkan perhitungan *Confusion Matrix* untuk menghitung akurasi algoritma *Random Decision Forest* dalam pengklasifikasian *file malware* PDF?

3. Bagaimana membandingkan tingkat keakuratan antara kedua metode yang dipakai?

1.3. Tujuan Penelitian

Berdasarkan permasalahan yang telah dijelaskan diatas, dapat disimpulkan tujuan dari penelitian ini adalah bagaimana pengklasifikasian *file* PDF yang terjangkit *malware* dan yang tidak, serta untuk mencari tahu seberapa akurat atara kedua algoritma yang dipakai yaitu *Support Vector Machine* dan *Random Forest.*

1.4. Batasan Masalah

Dalam penelitian ini perlu diberikan batasan masalah dengan tujuan agar pembahasan tidak meluas dan menyimpang dari tujuan. Adapun batasan masalah dalam penelitian ini antara lain:

1. Analisa *malware* hanya menggunakan format *file* PDF.
2. Sampel *malware* dan *non malware* didapat dari *Payload Security* (www.hybrid-analysis.com).
3. Untuk sampel tambahan *non malware* didapat dari penyimpanan pribadi.
4. Sampel *malware* hanya menggunakan 500 sampel yang terdiri dari data *malware* dan *non malware*.
5. Deteksi *malware* PDF hanya menggunakan teknik klasifikasi dengan metode *Support Vector Machine* dan *Random Decision Forest*.

1.5. Sistematika Penulisan

Sistematika penulisan laporan penelitian ini disusun menjadi beberapa bab sebagai berikut:

BAB I. PENDAHULUAN

Pada bab ini berisi pendahuluan yang menjelaskan latar belakang mengenai sebab dan pentingnya penelitian ini harus dilakukan, merumuskan

pokok permasalahan yang dihadapi, tujuan dilakukannya penelitian, batasan permasalahan, dan sistematika penulisan tugas akhir.

BAB II. LANDASAN TEORI

Pada bab ini membahas berbagai konsep dasar dan teori-teori yang berkaitan dengan topik penelitian tugas akhir ini dengan judul "Deteksi *Malware* Dalam *File Portable Document Format* (PDF) Menggunakan *Support Vector Machine* dan *Random Decision Forest*".

BAB III. ANALISA DAN PERANCANGAN SISTEM

Bab ini menjelaskan tentang rancangan sistem sesuai judul yang telah diajukan perancangan sisitem ini meliputi klasifikasi, pembobotan dan pembuatan aplikasi.

BAB IV. IMPLEMENTASI DAN PENGUJIAN SISTEM

Bab ini menjelaskan implementasi dari perancangan sistem yang telah dibuat pada bab III dan melakukan pengujian dari sistem tersebut. Berikut adalah skenario pengujiannya:

1. Perhitungan akurasi dari metode *Support Vector Machine* menggunakan *Confusion Matrix.* Sebagaimana yang dilakukan oleh penelitian sebelumnya yang menggunakan perhitungan *Confusion Matrix* untuk menghitung tingkat keakuratannya [11].
2. Perhitungan akurasi dari metode *Random Decision Forest* menggunakan *Confusion Matrix.* Sebagaimana yang dilakukan oleh penelitian sebelumnya yang menggunakan perhitungan *Confusion Matrix* untuk menghitung tingkat keakuratannya [10].
3. Membandingkan kedua hasil perhitungan *Confusion Matrix* dari kedua metode yang dipakai.

BAB V. PENUTUP

Pada bab ini berisi tentang kesimpulan dari penelitian yang telah dilakukan serta saran untuk pengembangan penelitian lebih lanjut.